

Sparse Principal Component Analysis: a Least Squares approximation approach

Giovanni Maria Merola*

Department of Economics, Finance and Marketing
RMIT International University Vietnam
702 Nguyen Van Linh
Dist 7. Ho Chi Minh City, Vietnam
(giovanni.merola@rmit.edu.vn)

August 19, 2014

Abstract

Sparse Principal Components Analysis aims to find principal components with few non-zero loadings. We derive such sparse solutions by adding a genuine sparsity requirement to the original Principal Components Analysis (PCA) objective function. This approach differs from others because it preserves PCA's original optimality: uncorrelatedness of the components and Least Squares approximation of the data. To identify the best subset of non-zero loadings we propose a Branch-and-Bound search and an iterative elimination algorithm. This last algorithm finds sparse solutions with large loadings and can be run without specifying the cardinality of the loadings and the number of components to compute in advance. We give thorough comparisons with the existing Sparse PCA methods and several examples on real datasets.

Keywords: SPCA; Uncorrelated Components; Branch-and-Bound; Backward Elimination
MSC-class: 62H12; 62H25

*We would like to thank Dr. Alessio Farcomeni for making available his R code for the branch-and-bound search and Professor Bob Baulch of RMIT Vietnam for his useful comments.

1 Introduction

Principal Component Analysis (PCA) is one of the most frequently used methods for approximating a set of variables with few linear combinations of them, called *principal components* (PCs) (e.g. Izenman, 2008). PCA was originally introduced by Pearson (1901) to find the “*lines and planes of closest fit to systems of points*” and was later rediscovered by Hotelling (1933), who wanted to find “*some more fundamental set of independent variables [...] which determine the values the x ’s will take.*” The PCs are estimated by minimising the sum of squares residuals of the approximation, in the words of Pearson (1901) “*a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line or planes a minimum.*” Hence, PCA belongs to the class of multivariate methods that optimise the Least Squares (LS) criterion to find the estimates (ten Berge, 1993).

The PCs are mutually uncorrelated and reproduce, or *explain*, the most possible variance of the data. Hotelling (1933) showed that the PCs are also the components with orthonormal loadings that sequentially have largest variance. PCA has been popularised in this simpler form often without mention of the original LS fit requirement.

The weights of the PCs (*loadings*) are used to interpret the PCs as meaningful characteristics of the data. For example, a component from a set of IQ test scores defined as $(\frac{1}{2} \text{ Inductive reasoning} + \frac{1}{2} \text{ Deductive reasoning})$ could be interpreted as the *Logic skills* of a person. However, since in most applications all the loadings are different from zero, convincing interpretations are difficult to find. The PCs would be easier to interpret if only few of the loadings were different from zero, that is, if they were *sparse*.

The most common way to achieve sparseness is by *thresholding* the loadings, that is, by discarding “small” ones (hereafter, for simplicity, we speak of the size of the loadings meaning the absolute value of the non-zero ones). Thresholding affects the optimality and uncorrelatedness of the solutions in an unpredictable way (Cadima and Jolliffe, 1995). Therefore, a number of *Sparse Principal Component Analysis* (SPCA) methods for estimating components with few non-zero loadings have been proposed. The number of non-zero loadings is called *cardinality* or L_0 norm.

The existing SPCA methods maximise the variance of the components under cardinality constraints, in most cases also requiring that the loadings are orthonormal. However, because of the additional constraints, this is no longer equivalent to maximising the variance explained. So, the components obtained simply maximise a numerical property of the PCs that is irrelevant for summarising the information contained in the data (ten Berge, 1993).

The lack of equivalence between objective function and variance explained created confusion in the SPCA literature on how components subsequent to the first one and the variance that they explain should be computed (e.g. Mackey, 2009 and Wang and Wu, 2012). Furthermore, in the absence of specific constraints, the SPCA components are correlated. Correlated components are more difficult to interpret and the sum of the variance that they explain individually is larger than the variance that they explain together.

Finding the optimal SPCA solutions is a nonconvex NP-hard, hence intractable, problem (Moghaddam et al., 2006). The existing SPCA methods use sophisticated numerical algorithms to approximately solve it. In this paper we will consider the estimation criteria used by these methods but not the numerical approximations proposed. A review of various SPCA methods can be found in Trendafilov (2013).

In this paper we improve on the existing methods by deriving uncorrelated sparse PCs that minimise the LS criterion. In this method, to which we refer as LS-SPCA, the solutions are obtained by adding cardinality restrictions to the PCA problem. The solutions can be computed as a series of constrained Reduced Rank Regression components (RRR, Izenman, 1975), for which we provide the closed form solutions.

Also for LS-SPCA the problem of finding the best set of loadings of a given cardinality

is intractable. As a first approach, for its solution we suggest a simple greedy iterative Branch-and-Bound (BB) search based on that proposed by Farcomeni (2009). We will show that it can be applied to moderately large size problems in reasonable time.

SPCA should replace manual thresholding, which produces interpretable sparse solutions with few big loadings. This is not guaranteed by the constraints on the L_0 norm and the sparse components computed may still be difficult to interpret. For this reason, we propose an algorithm that computes the sparse solutions by iteratively eliminating the smallest loadings. This greedy algorithm, to which we refer to as Backward Elimination (BE), at each iteration eliminates the smallest loading until only loadings larger than a threshold remain. Since the size of the loadings is not the only criterion to be considered, we also include rules for terminating the elimination when a maximum amount of variance explained is lost or a minimal cardinality is reached; these rules can be activated simultaneously.

The BE algorithm represents a departure from an optimal search. However, in terms of user's needs, the increased interpretability of the solutions should compensate the ensuing loss of variance explained. In the numerical section we show that the BE components compare well with the BB ones. Furthermore, with this flexible algorithm, cardinality, minimum size of the loadings and departure from the optimum can be controlled at the same time. These controls also eliminate the need for specifying in advance the cardinalities and the number of components to compute.

The selection of a sparse set of loadings is similar to model selection in Linear Regression, where there is a competition between parsimony of the model and variance of the response explained. In the same way, there exist similar sparse solutions that compete on variance explained and interpretability. The BE algorithm can be also used in an explanatory phase of the analysis for comparing different solutions.

The BB and BE algorithms are implemented in an R package, which will soon be released. In this implementation the sparse loadings can be restricted to include only a subset of variables and more than one loading can be eliminated when there are a large number of variables. The package also contains methods for producing summaries, plots and comparisons of the results. Some benchmark datasets are also included.

The original paper proposing LS-SPCA is still unpublished because reviewers seem to ignore that the original objective of PCA is the maximisation of the variance explained. Also journals editors seem to share the same lack of knowledge and to approve biased reviews, likely from authors of different SPCA methods. For example, Dr. Qiu, the chief editor of *Technometrics*, accepted a report which did not discuss at all the content but rejected the paper because I would ignore an article (by A. D'asprenmont *et al.*) which, instead, was cited in the references and the results of which were compared with mine (as in this version). He also accepted another report from a referee (who could barely write in English) who called the LS criterion "a new measure used ad-hoc". Dr Qiu rejected the paper on those grounds and added the reason that the algorithm is not scalable. Since computational efficiency is not in the scope of *Technometrics*, this also was unfair. The (kind) letter of complaint I sent to Dr. Qiu went unanswered. This is just to testify how frustrating publishing sometimes can be, thanks to unfair resistance from other fellow academics.

The rest of the paper is organized as follows: in the next section we formalise the SPCA problem into the LS optimisation framework and comparing it with the estimation criteria used in other SPCA methods. In Section 3 we derive the closed form solutions for LS-SPCA method, also considering correlated ones. In Section 4 we present the branch-and-bound search and the Backward Elimination algorithm. In Section 5 we give extensive numerical comparisons with other SPCA methods on benchmark datasets and show the results of our methods on real life datasets of varying dimensions. Finally, in Section 6 we give some concluding remarks.

2 Sparse Principal Components Analysis

Statistical estimation theory requires that a measure of goodness of fit is defined and optimised. Pearson (1901) explicitly adopts the Least Square criterion. Hotellings (1933) paper is quite intricate and the estimation procedure used is not clearly stated. The solutions are the same as Pearsons ones but are given as the components with orthonormal (that is with unit Euclidean norm (L2) and mutually orthogonal) loadings that have maximum variance. Due to its simplicity, this has become the standard definition with which PCA has been popularised among practitioners. It must be noted that Pearsons solutions do not have the properties that he was looking for. In fact, in the introduction of his paper (page 417) he stated *We shall consider only normally distributed systems of components having zero correlations and unit variances*, and the last two properties are paramount in his derivation. Obviously, the maximal variance of the components cannot be constantly equal to one and orthogonality of the loadings is not sufficient for uncorrelatedness. Hence, the estimation did not solve the simplified problem.

In this section we first give an overview of the LS derivation of the PCs, because it is less known than others. Then we define the sparse PCA estimation criterion, or problem, by adding L_0 constraints to the LS optimisation. Finally, we discuss the estimation criteria adopted by other SPCA methods.

2.1 Notation

In the following, we denote matrices with bold uppercase letters, \mathbf{X} , and vectors with bold lowercase ones \mathbf{a} . The columns of a matrix are denoted with the corresponding bold lowercase symbol, indexed accordingly, \mathbf{x}_j . The symbol \mathbf{I}_k denotes the identity matrix of order k . The reference to the order is omitted when this is clear from the context. $\text{tr}(\mathbf{S}) = \sum_{j=1}^p s_{j,j}$ denotes the trace of a $(p \times p)$ square matrix. The L_2 norm will be denoted as $\|\cdot\|$, so that $\|\mathbf{X}\| = [\text{tr}(\mathbf{X}'\mathbf{X})]^{\frac{1}{2}}$. A "hat" on a symbol denotes an estimate and the subscript $[\cdot]$ denotes the rank of a matrix, so, for example, $\hat{\mathbf{X}}_{[d]}$ is the estimate of a rank- d approximation of the data matrix \mathbf{X} . A "*" denotes the globally optimal PCA estimates. The subscript (d) denotes the first $(d-1)$ columns of a matrix, so that $\mathbf{A}_{(j)}$ are the first $(j-1)$ columns of \mathbf{A} .

2.2 The PCA problem

Given a matrix of n observations on p mean-centred variables \mathbf{X} ($n \times p$), PCA intends to find rank- d ($d \leq p$) approximation of the data, $\hat{\mathbf{X}}_{[d]}$. It is easy to prove that the approximation can be written as $\hat{\mathbf{X}}_{[d]} = \mathbf{X}\mathbf{A}\mathbf{P}'$, where \mathbf{A} ($p \times d$) is the matrix of loadings, $\mathbf{T} = \mathbf{X}\mathbf{A}$ ($n \times d$) the matrix of the PCs and \mathbf{P} ($d \times p$) a matrix of coefficients. The solutions are determined by minimising the LS criterion, that is, as:

$$\arg \min_{\text{Rank}(\hat{\mathbf{X}}_{[d]})=d} \|\mathbf{X} - \hat{\mathbf{X}}_{[d]}\|^2 = \arg \min_{\mathbf{A} \in \mathbb{R}^{p \times d}} \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{P}'\|^2$$

The solutions are completely identified by the loadings because $\mathbf{P}' = (\mathbf{T}'\mathbf{T})^{-1} \mathbf{T}'\mathbf{X}$ so that the rank- d approximation is equal to $\hat{\mathbf{X}}_{[d]} = \mathbf{X}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A})^{-1} \mathbf{A}'\mathbf{X}'\mathbf{X}$.

The components can be constrained to be uncorrelated without loss of optimality. In fact, by the principle of the *extra sum of squares* a set of correlated components cannot explain more variance than the same number of uncorrelated ones (in the appendix we provide a proof of this well known result). Beside ease of interpretation, another advantage of uncorrelatedness is that the resulting ordered PCs are the LS estimates for any number of components included in the model.

If we let $\mathbf{S} = n^{-1}\mathbf{X}'\mathbf{X}$ ($p \times p$) denote the sample covariance matrix, under uncorrelatedness constraints the PCA problem becomes:

$$\begin{aligned} \mathbf{A} = & \arg \min_{\text{Rank}(\hat{\mathbf{X}}_{[d]}=d)} \|\mathbf{X} - \hat{\mathbf{X}}_{[d]}\|^2 = \arg \max_{\mathbf{A} \in \mathbb{R}^{p \times d}} \|\hat{\mathbf{X}}_{[d]}\|^2 = \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \sum_{j=1}^d \frac{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j} \quad (1) \\ & \text{subject to } \mathbf{a}_j' \mathbf{S} \mathbf{a}_k = 0, \quad j \neq k, \end{aligned}$$

where the summation in the last term derives from the uncorrelatedness of the components. It follows that the total variance explained can be broken down into the sum of the variances explained by each component. Therefore the LS criterion is equivalent to the maximisation of the variances explained by each component, $\text{Vexp}(\mathbf{t}_j)$. Hence, the PCA problem can be equivalently written as:

$$\begin{aligned} \mathbf{a}_j = & \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \text{Vexp}(\mathbf{t}_j) = \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \frac{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}. \quad (2) \\ & \text{subject to } \mathbf{b}_j \mathbf{S}' \mathbf{a}_k = \delta_{jk} \end{aligned}$$

2.3 The PCA solutions

If no other constraints are added to Problem (1), the PCA loadings are proportional to the eigenvectors of \mathbf{S} , $\{\mathbf{v}_j, j = 1, \dots, d\}$, corresponding to the d largest eigenvalues taken in non increasing order, $\{\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d\}$. Then it follows that the loadings are orthonormal and the variance explained by each PC is equal to the corresponding eigenvalue, because Equation (2) simplifies to:

$$\text{Vexp}^*(\mathbf{t}_j) = \frac{\mathbf{v}_j' \mathbf{S} \mathbf{v}_j}{\mathbf{v}_j' \mathbf{S} \mathbf{v}_j} = \frac{\mathbf{v}_j' \mathbf{S} \mathbf{v}_j}{\mathbf{v}_j' \mathbf{v}_j} = \lambda_j.$$

Hence the variance explained by a PC is equal to its variance. This property has led to the popularisation of PCA simply as the method that finds the components with orthonormal coefficients that have sequentially maximal variance. Consequently, the PCA problem is often defined as:

$$\begin{aligned} \mathbf{a}_j = & \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \mathbf{a}_j' \mathbf{S} \mathbf{a}_j, \quad j = 1, \dots, d \quad (3) \\ & \text{subject to } \mathbf{a}_j' \mathbf{a}_k = \delta_{jk}, \end{aligned}$$

where δ_{jk} is the Kronecker delta. ten Berge (1993, page 87) warns about this formulation of the PCA problem by stating: "*Nevertheless, it is undesirable to maximize the variance of the components rather than the variance explained by the components, because only the latter is relevant for the purpose of finding components that summarize the information contained in the variables.*"

2.4 The Least Squares Sparse PCA problem

The LS-SPCA Problem is obtained by constraining the cardinality of the loadings in Problem (1), which gives:

$$\begin{aligned} \mathbf{A} = & \arg \min_{\mathbf{A} \in \mathbb{R}^{p \times d}} \|\mathbf{X} - \mathbf{X} \mathbf{A} \mathbf{P}'\|^2 = \arg \max_{\mathbf{A} \in \mathbb{R}^{p \times d}} \sum_{j=1}^d \frac{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j} = \arg \max_{\mathbf{A} \in \mathbb{R}^{p \times d}} \sum_{j=1}^d \text{Vexp}(\mathbf{t}_j) \quad (4) \\ & \text{subject to } L_0(\mathbf{a}_j) \leq c_j \text{ and } \mathbf{a}_j' \mathbf{S} \mathbf{a}_k = 0, \quad j \neq k, \end{aligned}$$

where $c_j < p$ are the maximal cardinalities allowed.

As we will show in Section 3, under the cardinality constraints the loadings $\text{Vexp}(\mathbf{t}_j)$ no longer simplifies to $(\mathbf{a}_j' \mathbf{S} \mathbf{a}_j)/\mathbf{a}_j' \mathbf{a}_j$ and the solutions must be obtained by maximising Vexp in Equation (2) directly.

2.5 Other SPCA problems

In the existing SPCA methods the components are computed by maximising the variance of the components under cardinality constraints. Hence, the sparse PCA problem is defined by adding cardinality constraints to the simplified PCA problem (3), which gives:

$$\begin{aligned} \mathbf{b}_j &= \arg \max_{\mathbf{b}_j \in \mathbb{R}^p} \mathbf{b}_j' \mathbf{S} \mathbf{b}_j, j = 1, \dots, d \\ \text{subject to } \mathbf{b}_j' \mathbf{b}_k &= \delta_{jk} \text{ and } L_0(\mathbf{b}_j) \leq c_j, \end{aligned} \quad (5)$$

for some parameters c_j ($c_j < p$). Clearly, this problem is not analogous to Problem (4) because it implicitly assumes that the solutions are eigenvectors of \mathbf{S} , which they cannot be under sparsity requirements.

The SPCA solutions of a given cardinality c_j are the first eigenvectors of the $(c_j \times c_j)$ principal submatrix of \mathbf{S} with the largest maximal eigenvalue, subject to the constraints (Moghaddam et al., 2006, Proposition 1). Hence the SPCA problem can also be written in terms of the non-zero loadings $\tilde{\mathbf{b}}_j$ as

$$\begin{aligned} \tilde{\mathbf{b}}_j &= \arg \max_{\tilde{\mathbf{b}}_j \in \mathbb{R}^{c_j}} \tilde{\mathbf{b}}_j' \mathbf{D}_j \tilde{\mathbf{b}}_j, j = 1, \dots, d \\ \text{subject to } \tilde{\mathbf{b}}_j' \tilde{\mathbf{b}}_k &= \delta_{jk}, \end{aligned}$$

where \mathbf{D}_j is the principal submatrix of \mathbf{S} corresponding to indices of the sparse loadings. Hence, the SPCA problem boils down to finding the sets of indices, ind_j , that give the principal submatrix of \mathbf{S} with largest maximum eigenvalue.

In some SPCA methods (*e.g.* Moghaddam et al., 2006) the orthogonality constraints are omitted from the problem. In this case, trivial solutions are avoided by computing the solutions subsequent to the first one on the covariance matrix *deflated* in different ways. In general, there is no agreement on which is the correct deflation to use (see Mackey, 2009, and Wang and Wu, 2012, for a discussion). It should be noted that, the orthogonality constraints ensure that a full set of sparse components explains all of the data variance, while this is not guaranteed if the components are not uncorrelated and the loadings not orthogonal. We observe that a necessary condition for orthogonality is that the cardinality of the loadings is not smaller than their rank. This condition, as we will show in the next section, applies also to the uncorrelatedness constraints. The solutions to Problem (5) subsequent to the first one are given by

$$\begin{aligned} \mathbf{b}_j &= \arg \max_{\mathbf{b}_j \in \mathbb{R}^p} \mathbf{b}_j' \left(\mathbf{I} - \mathbf{B}_{(j)} (\mathbf{B}_{(j)}' \mathbf{B}_{(j)})^{-1} \mathbf{B}_{(j)}' \right) \mathbf{S} \left(\mathbf{I} - \mathbf{B}_{(j)} (\mathbf{B}_{(j)}' \mathbf{B}_{(j)})^{-1} \mathbf{B}_{(j)}' \right) \mathbf{b}_j \\ \text{subject to } \mathbf{b}_j' \mathbf{b}_j &= 1 \text{ and } L_0(\mathbf{b}_j) \leq c_j. \end{aligned}$$

These solutions correspond to the *additional variance* deflation, derived by Mackey (2009) from different properties.

Some authors assume without justification that the variance of the sparse components is equal to the variance explained (*e.g.* d'Aspremont et al., 2008). In other cases the maximisation of the variance of the components is derived from different problems. Zou et al. (2006) adopt a LS optimisation subject to L_1 constraints, in a Lasso fashion. However, they also constrain the coefficients \mathbf{p}_j to have unit variance, that is, they require that $\|\mathbf{p}_j\|^2 = 1$. Since $\mathbf{p}_j = \mathbf{S} \mathbf{b}_j / (\mathbf{b}_j' \mathbf{S} \mathbf{b}_j)$, these constraints are equivalent to requiring that

$$\frac{\mathbf{b}_j' \mathbf{S} \mathbf{S} \mathbf{b}_j}{\mathbf{b}_j' \mathbf{S} \mathbf{b}_j} = \frac{\mathbf{b}_j' \mathbf{S} \mathbf{b}_j}{\mathbf{b}_j' \mathbf{b}_j},$$

where the right hand side is equivalent to the variance of the components with normal loadings. In this way, the scope of the objective is limited because, by the Cauchy-Schwartz inequality:

$$\frac{\mathbf{b}_j' \mathbf{S} \mathbf{S} \mathbf{b}_j}{\mathbf{b}_j' \mathbf{S} \mathbf{b}_j} \geq \frac{\mathbf{b}_j' \mathbf{S} \mathbf{b}_j}{\mathbf{b}_j' \mathbf{b}_j},$$

for any square matrix \mathbf{S} . It should be noted that Zou et al. (2006) want to achieve Lasso type "shrinkage" estimates and not traditional LS ones.

Shen and Huang (2008) justify the maximisation of the variance of the components considering that the PCA solutions can be derived from the maximisation of the trace of the approximation of the covariance matrix. The approximation is $\hat{\mathbf{S}}_{[d]}^* = \hat{\mathbf{X}}_{[d]}^* \hat{\mathbf{X}}_{[d]}^* = \sum_{j=1}^d \mathbf{v}_j \mathbf{v}_j' \lambda_j$, with $\lambda_j = \mathbf{v}_j' \mathbf{S} \mathbf{v}_j$ and $\mathbf{v}_j' \mathbf{v}_k = \delta_{jk}$. Therefore, they seek the orthonormal sparse vectors \mathbf{b}_j with maximal variance so that the approximate covariance matrix is $\mathbf{S}_{[d]} = \sum_{j=1}^d \mathbf{b}_j \mathbf{b}_j' (\mathbf{b}_j' \mathbf{S} \mathbf{b}_j)$. However, also in this case, the LS-SPCA loadings, \mathbf{a}_j dominate the SPCA solutions. In fact, if we define the orthonormal vectors $\mathbf{e}_j = \mathbf{S}^{\frac{1}{2}} \mathbf{a}_j / (\mathbf{a}_j' \mathbf{S} \mathbf{a}_j)^{\frac{1}{2}}$, we can write the LS-SPCA approximation as $\hat{\mathbf{S}}_{[d]} = \sum_{j=1}^d \mathbf{e}_j \mathbf{e}_j' \rho_j$ with $\rho_j = \mathbf{e}_j' \mathbf{S} \mathbf{e}_j = \mathbf{a}_j' \mathbf{S} \mathbf{S} \mathbf{a}_j / (\mathbf{a}_j' \mathbf{S} \mathbf{a}_j)$. Then,

$$\text{tr}(\hat{\mathbf{S}}_{[d]}) = \sum_{j=1}^d \frac{\mathbf{a}_j' \mathbf{S} \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j} \geq \sum_{j=1}^d \frac{\mathbf{b}_j' \mathbf{S} \mathbf{S} \mathbf{b}_j}{\mathbf{b}_j' \mathbf{S} \mathbf{b}_j} \geq \sum_{j=1}^d \frac{\mathbf{b}_j' \mathbf{S} \mathbf{b}_j}{\mathbf{b}_j' \mathbf{b}_j} = \text{tr}(\mathbf{S}_{[d]}),$$

where \mathbf{a}_j and \mathbf{b}_j have the same cardinality. Even not considering this result, it is still unexplained in which sense the resulting vectors \mathbf{b}_j could be used as loadings for the components, because in this case the approximation of the variance would be $\mathbf{S}_{[d]} = \mathbf{S} \mathbf{B} (\mathbf{B} \mathbf{S} \mathbf{B})^{-1} \mathbf{B} \mathbf{S}$ and not the one optimised. Furthermore, when the components are correlated the trace of the approximation is no longer the sum of the individual approximations but the sum of the diagonal elements of the matrix $\mathbf{S}_{[d]}$.

3 Least Squares Sparse PCA solutions

If we assume that the indices of the sparse loadings of the d required components, ind_j , are known, the sparse components will be combinations of only the corresponding variables, denoted with the matrices \mathbf{W}_j ($n \times c_j$). Then, the sparse components can be written as $\mathbf{t}_j = \mathbf{W}_j \tilde{\mathbf{a}}_j$, where the $\tilde{\mathbf{a}}_j$ are the vectors of dimension c_j containing only the non-zero loadings. With this notation we can write the individual LS-SPCA problems as:

$$\tilde{\mathbf{a}}_j = \arg \min_{\tilde{\mathbf{a}}_j \in \mathbb{R}^{c_j}} \|\mathbf{X} - \mathbf{W}_j \tilde{\mathbf{a}}_j \mathbf{p}_j'\|, \quad j = 1, \dots, d \quad (6)$$

$$\text{subject to } \mathbf{T}_{(j)}' \mathbf{W}_j \tilde{\mathbf{a}}_j = \mathbf{0}, \quad j > 1, \quad (7)$$

where $\mathbf{T}_{(j)}$ is the matrix containing the first $(j-1)$ components ($\mathbf{T}_{(1)} = \mathbf{0}$). Let \mathbf{J}_j ($p \times c_j$) be the matrices formed by the columns of the p dimensional identity matrix with indices in ind_j , then we can write $\mathbf{W}_j = \mathbf{X} \mathbf{J}_j$ and the full sparse loadings as $\mathbf{a}_j = \mathbf{J}_j \tilde{\mathbf{a}}_j$. With this notation, the SPCA problem defined in Equation (6) can be written as

$$\arg \min_{\tilde{\mathbf{a}}_j \in \mathbb{R}^{c_j}} \|\mathbf{X} - \mathbf{W}_j \tilde{\mathbf{a}}_j \mathbf{p}_j'\| = \arg \max_{\tilde{\mathbf{a}}_j \in \mathbb{R}^{c_j}} \frac{\tilde{\mathbf{a}}_j' \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j}{\tilde{\mathbf{a}}_j' \mathbf{J}_j' \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j} = \arg \max_{\tilde{\mathbf{a}}_j \in \mathbb{R}^{c_j}} \frac{\mathbf{a}_j' \mathbf{S} \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}, \quad j = 1, \dots, d \quad (8)$$

subject to $\mathbf{R}_j \mathbf{a}_j = \mathbf{0}$, for $j > 1$,

where $\mathbf{R}_j = \mathbf{A}'_{(j)} \mathbf{S} \mathbf{J}_j$ defines the uncorrelatedness constraints, with $\mathbf{A}_{(j)}$ being the first $(j - 1)$ loadings. Hence the sparse PCs maximise Vexp defined in Equation (2), under the constraints.

Problem (8) can be seen as a series of constrained rank-one Reduced-Rank Regression problems where the regressors are the columns of the \mathbf{W}_j matrices. It is well known that the first solution is the eigenvector $\tilde{\mathbf{a}}_1$ satisfying:

$$(\mathbf{W}'_1 \mathbf{W}_1)^{-1} \mathbf{W}'_1 \mathbf{X} \mathbf{X}' \mathbf{W}_1 \tilde{\mathbf{a}}_1 = \phi_{\max} \tilde{\mathbf{a}}_1, \quad (9)$$

where ϕ_{\max} is the largest eigenvalue. This solution is unique as long as the variables in \mathbf{W}_1 are not multicollinear. Hereafter we exclude the possibility that a matrix \mathbf{W}_j is not full column rank because that set of variables should be discarded and a full rank one sought. The sparse loadings can be computed also if only the covariance matrix \mathbf{S} is known. In fact, Equation (9) can be written as:

$$\mathbf{D}_1^{-1} \mathbf{J}'_1 \mathbf{S} \mathbf{J}_1 \tilde{\mathbf{a}}_1 = \phi_{\max} \tilde{\mathbf{a}}_1, \quad (10)$$

where $\mathbf{D}_j = \mathbf{W}'_j \mathbf{W}_j = \mathbf{J}'_j \mathbf{S} \mathbf{J}_j$ is the covariance matrix of the variables with index in ind_j , which is invertible for the full rank assumptions.

The following solutions can be found by applying the uncorrelatedness constraints to the RRR Problems (8) as in constrained multiple regression (*e.g.*, see Rao and Toutenburg, 1999, or Magnus and Neudecker, 1999, Th. 13.5, for a more rigorous proof). In the appendix we show that these are given by the eigenvectors satisfying

$$\mathbf{C}_j \mathbf{D}_j^{-1} \mathbf{J}'_j \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j = \phi_{\max} \tilde{\mathbf{a}}_j, \quad (11)$$

where $\mathbf{C}_j = \mathbf{I}_{c_j} - \mathbf{D}_j^{-1} \mathbf{R}'_j (\mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{R}'_j)^+ \mathbf{R}_j$, with $\mathbf{C}_1 = \mathbf{I}_{c_1}$, and the subscript "+" denotes a generalized inverse. The solutions exist because \mathbf{R}_j spans the space of \mathbf{W}_j . In this derivation we assume that $\mathbf{R}'_j \mathbf{R}_j$ is singular, otherwise $\mathbf{R}_j \tilde{\mathbf{a}}_j = \mathbf{0}$ can never be satisfied. This means that uncorrelatedness can only be achieved if the cardinalities satisfy $c_j \geq j$. The LS-SPCA solutions can be computed from the leftmost eigenvector, \mathbf{b}_j , of the symmetric matrices $(\mathbf{C}_j \mathbf{D}_j^{-1})^{\frac{1}{2}} \mathbf{J}'_j \mathbf{S} \mathbf{J}_j (\mathbf{C}_j \mathbf{D}_j^{-1})^{\frac{1}{2}}$ as $\tilde{\mathbf{a}}_j = (\mathbf{C}_j \mathbf{D}_j^{-1})^{\frac{1}{2}} \mathbf{b}_j$.

The above derivation shows that the sparse components that explain the most variance are not eigenvectors of submatrices of the covariance matrix and that their variance is no longer equal to the variance that they explain.

As mentioned above, in LS-SPCA the uncorrelatedness constraints require that the cardinality of the components is not less than their order. Correlated component of lower cardinality can be computed by applying LS-SPCA to the residual of \mathbf{X} orthogonal to the previous components, $\mathbf{X}_j = \mathbf{I} - \mathbf{T}_{(j)} (\mathbf{T}'_{(j)} \mathbf{T}_{(j)})^{-1} \mathbf{T}'_{(j)} \mathbf{X}$. While correctly used in iterative PCA algorithms, such as Power Method and NIPALS (Wold, 1966), for example, under sparsity constraints this approach does not maximize Vexp in Equation (2), but the approximations

$$\frac{\mathbf{a}'_j \mathbf{S}_j \mathbf{S}_j \mathbf{a}_j}{\mathbf{a}'_j \mathbf{S} \mathbf{a}_j}, \quad (12)$$

hence the suboptimality.

In the appendix we show that these correlated components are given by the eigenvectors satisfying

$$\mathbf{D}_j^{-1} \mathbf{J}'_j \mathbf{S}_j \mathbf{S}_j \tilde{\mathbf{a}}_j = \phi_{\max} \tilde{\mathbf{a}}_j, \quad (13)$$

where the \mathbf{S}_j are the covariance matrices computed from the residuals \mathbf{X}_j . When needed, we will refer to these solutions as Least Squares Correlated Sparse Principal Components Analysis (LS-CSPCA).

Following the same approach, correlated sparse components with orthogonal loadings can also be found. In the appendix we show that these solutions are the eigenvectors satisfying:

$$\mathbf{D}_j^{-1} \mathbf{J}_j' \left[\mathbf{I}_p - \mathbf{A}_{(j)} \left(\mathbf{A}_{(j)}' \mathbf{J}_j \mathbf{D}_j^{-1} \mathbf{J}_j' \mathbf{A}_{(j)} \right)^{-1} \mathbf{A}_{(j)}' \mathbf{J}_j \mathbf{D}_j^{-1} \mathbf{J}_j' \right] \mathbf{S}_j \mathbf{S}_j' \mathbf{J}_j \tilde{\mathbf{a}}_j = \phi_{max} \tilde{\mathbf{a}}_j.$$

4 Determining the indices of the sparse loadings

Solving the LS-SPCA problem requires determining the optimal set of indices for d components. The cardinality constraints make it a non-convex problem, which cannot be efficiently solved. We consider first a greedy Branch-and-Bound search (BB) then we consider a greedy Backward Elimination algorithm designed to replace manual thresholding.

4.1 LS-SPCA(BB): a branch-and-bound search for the optimal loadings

If the cardinalities of each component are specified, locally optimal subsets of loadings can be found through a greedy BB search based on that proposed by Farcomeni (2009) using Vexp in Equations (2) as bounding function. At each node a subsets of variables of cardinality greater than the required one is discarded if the solution explains less variance than the current upper bound. The search continues on to subsets of smaller size until the best set of the required cardinality is found. This procedure leads to the optimal solution for each component because eliminating a variable from a regression model cannot yield a higher regression sum of squares. For correlated components, we use the variance explained by the approximate solutions, Equation (12), because the true variance explained, Equation (2), is not monotonically larger.

The search can start from a subset of indices instead of the complete one, if the analyst has a tentative solution in mind. The algorithm can be speeded up by sorting the variables with respect to the variance of the residuals that they individually explain. Note that the solutions are only locally optimal for each component, because the search does not explore all combinations of loadings for the given number of components. We will refer to the solutions obtained using this BB algorithm as LS-SPCA(BB) and LS-CSPCA(BB) for the uncorrelated and correlated solutions, respectively.

LS-SPCA(BB) is not computationally efficient and can only be run on medium or small size problems within reasonable time. Farcomeni (2009) gives an account of the computational times taken by the BB search for his method; in LS-SPCA(BB) it would take longer because the solutions are more complex to compute. In spite of this, the BB can be run on moderate size problems which are typically the ones in which the loadings are interpreted.

4.2 LS-SPCA(BE): a backward elimination for thresholding the loadings

Like all other SPCA methods, LS-SPCA(BB) has the drawback that some of the sparse loadings computed may be small, making the solutions still difficult to interpret. As a matter of fact, SPCA should replace thresholding, which gives only big sparse loadings. Hence, we consider the problem of determining the sparse solutions so that the sparse loadings are larger than a given threshold value. Let the thresholds for each dimension be denoted as τ_j ($0 < \tau_j \leq 1$), then the problem can be formalized as

$$\begin{aligned} \mathbf{a}_j &= \arg \min_{\mathbf{a}_j \in \mathbb{R}^p} \|\mathbf{X} - \mathbf{X} \mathbf{a}_j \mathbf{p}_j'\|, \quad j = 1, \dots, d \\ &\text{subject to } \left\{ a_{ij} = 0 \text{ or } \frac{|a_{ij}|}{L.(\mathbf{a}_j)} > \tau_j \right\} \text{ and } \mathbf{t}_j' \mathbf{t}_k = 0 \quad j \neq k. \end{aligned} \tag{14}$$

where $L(\mathbf{a}_j)$ denotes a norm, typically the L_1 or L_2 norm. This problem is NP-hard and BB type searches are difficult to derive because bounding functions are not obvious. However, if the variables are standardized to the same length, it can be expected that eliminating a small loading will not decrease V_{exp} by much. Therefore, we suggest a simple greedy backward elimination algorithm to iteratively eliminate the smallest sparse loadings from a solution until only ones larger than a given threshold are left. In this procedure it is not necessary to specify the cardinality of the solutions in advance and the elimination can be stopped by criteria different from the minimum threshold. For example, if only j non-zero loadings are left, the elimination must be stopped to maintain uncorrelatedness.

Wang and Wu (2012) proposed a backward elimination algorithm for SPCA based on different criteria. They mention examples in which eliminating small loadings is highly unreliable for SPCA. We believe that this is true when the components are correlated and the loadings are computed as pseudo-eigenvectors of a deflated matrix. In our studies we found our BE procedure very reliable and, in some cases, better than the BB search, as will be shown in the examples in Section 5.

The BE algorithm is outlined in Algorithm 1.

Algorithm 1 LS-SPCA(BE)

```

initialize
  Stopping rules for the number of components
     $d$  {the number of components to compute}
     $mv$  {optional, minimum variance cumulated explained for ending the algorithm}
  Stopping rules for elimination. Can be different for each component
     $ind_j$  {starting set of indices}
     $\tau_j$  {minimum absolute value of the sparse loadings}
     $k_j \geq 1$  {minimum cardinality of the sparse loadings}
     $mvl_j$  {optional maximum relative loss of variance explained}
end initialize
for  $j = 1$  to  $d$  do
  Compute  $\mathbf{a}_j$  as the  $j$ -th LS-SPCA solution for  $ind_j$ 
   $V_{\text{expfull}} = V_{\text{exp}}(\mathbf{a}_j)$ 
  while  $\min_{i \in ind_j} |a_{ij}| > \tau_j$  and  $length(ind_j) > k_j$  do
     $indold_j = ind_j$ ,  $\mathbf{aold}_j = \mathbf{a}_j$ 
     $k$ :  $|a_{kj}| \leq |a_{ij}|, i \in ind_j$ 
     $ind_j = ind_j \setminus k$ 
    Compute  $\mathbf{a}_j$  as the  $j$ -th LS-SPCA solution for  $ind_j$ 
    if  $1 - V_{\text{exp}}(\mathbf{a}_j)/V_{\text{expfull}} > mvl_j$  then
       $ind_j = indold_j$ ,  $\mathbf{a}_j = \mathbf{aold}_j$ 
      break
    end if
  end while
  if  $\sum_{i=1}^j V_{\text{exp}}(\mathbf{a}_i) \geq mv$  then
     $d = j$ 
    break
  end if
end for

```

Depending on the choice of the thresholds, trimming may cause too large a loss of V_{exp} or yield solutions with too few non-zero loadings. Therefore, we include in the algorithm two optional additional stopping rules for trimming: one based on the required minimum cardinality (which is needed if uncorrelated components are sought) and the other based on the maximal loss of V_{exp} induced by the last trimming. These stopping criteria can

be used instead of specifying the thresholds or additionally to it. Specifying in advance the number of components to compute can be difficult. So, we include also an optional rule for stopping the algorithm when a specified proportion of total variance explained is reached. This rule can be used together with a specified maximum number of components to be computed. The basic LS-SPCA Backward Elimination algorithm (LS-SPCA(BE)) is outlined in Algorithm 1:

Hence, the LS-SPCA(BE) algorithm can be used as a flexible tool without the need of specifying in advance the number of components to compute and their cardinality. In order to decrease the computational time for large matrices, more than one loading can be trimmed at each iteration. In this case, better results can be reached if, instead of terminating trimming when the cardinality is smaller than the number of loadings to be trimmed, the process is continued by trimming the remaining loadings individually until a stop rule is reached. Trimming can also be started from a subset of indices, if there is a reason for excluding some of the variables from a component.

For uncorrelated components, the minimal cardinalities must be set so as $k_j \geq j$, otherwise uncorrelatedness cannot be achieved (unless the reduced set is multicollinear). In this case, solutions fully trimmed to a given threshold may not be obtained. The stoprule for ending trimming can be defined with respect to different criteria, for example the loss of Vexp from the initial solution (as in Algorithm 1) or the loss of Cumulative Vexp from the corresponding PCA value. There is no obvious rule for choosing the thresholds τ_j . However, if the loadings are computed to have unitary L_2 norm, setting $\tau_j > 1/\sqrt{c}$ will ensure a cardinality lower than c (almost surely). In some cases, the size of the loadings is easier to evaluate if they are expressed as percentage *contributions*, that is, are standardized to unit L_1 norm. In this case, setting $\tau_j > 1/c$ will ensure a cardinality lower than c . For this reason, the choice of the minimum cardinality and of the threshold must be considered together and later components require a lower threshold than the first ones. The minimal total variance to be explained for ending the algorithm can be chosen with respect to the Vexp explained by the PCs. Note that trimming is designed for components computed from correlation matrices. If a covariance matrix is used, different thresholds for every variable should be used.

5 Numerical comparisons and examples

In this section we first compare LS-SPCA with other existing SPCA methods on two benchmark datasets. Then we show the results of LS-SPCA on several publicly available datasets. All the examples were computed on correlation matrices but we refer to the variance explained rather than to the correlation explained for uniformity with the literature. In order to make fair comparisons, the variance explained was recomputed according to Equation (2) for all the solutions from other methods. Loadings smaller than 0.001 in absolute value are not shown.

5.1 Comparison with other methods

In this section we compare the LS-SPCA results with those of other SPCA methods published for the only two benchmark datasets available in the literature, the famous Pitprop dataset (Thurstone, 1947) and an artificial one proposed by Zou et al. (2006). Therefore, the list of methods included is necessarily not exhaustive but it contains all the methods which, to our knowledge, were tested on these datasets. Unfortunately, the datasets are small in size, therefore differences in the variance explained are also small. Furthermore, the solutions in most cases were obtained for different cardinalities. This means that we can only compare LS-SPCA with each one of them and comparisons between different SPCA methods are not possible.

In our comparison we include some methods that relax the cardinality constraints by replacing it with constraints in the L_1 norm of the loadings. These are: SCoTLAS (Jolliffe et al., 2003; Trendafilov and Jolliffe, 2006); *SPCA* (Zou et al., 2006) who used a Lasso approach, and d’Aspremont et al. (2007), DSPCA. Other methods that tackle the L_0 constraints are: Moghaddam et al. (2006)’s greedy and exact Branch-and-Bound algorithms, GSPCA and ESPCA, respectively; Sriperumbudur et al. (2009), DCLS-SPCA, Shen and Huang (2008), sPCA-rSVD, and Journée et al. (2010), *Gpower*; Farcomeni (2009), BB-sPCA; Wang and Wu (2012) who proposed a simple but effective greedy backward elimination algorithm, SPCA-IE, for the same problem. Lastly, Trendafilov (2013) offers a number of solutions based on different approximations of the covariance matrix under L_1 constraints.

It should be noted that the comparisons are useful only to gain a feel of how different SPCA solutions compare to the LS-SPCA BB and BE algorithms. In fact, we showed that theoretically the LS-SPCA solutions are the ones that maximise the variance explained but the greedy algorithms may not find the global optima. It should also be considered that the BE algorithm is designed to attain large loadings at the cost of explaining less variance. Furthermore, our algorithms are implemented in R without optimising the code. Therefore scalability or speed issues will not be addressed. In spite of this we were able to compute sparse solutions for problems with up to almost 1000 variables in few minutes on a small computer.

5.1.1 Performance measures

Obviously, a performance measure of a set of SPCA solutions is the variance explained by each component, computed according to Equation (2), and the cumulated one. Hence we report the percentage of individual and cumulated variance explained, denoted as PVE and PCVE, respectively. The variance explained by the full PCs is an upper bound for the variance explained by the sparse components, hence we also consider the cumulated variance explained relative to that of PCA, denoted as PRCVE. A crucial feature of the sparse solutions is the cardinality of the loadings, we report these denoted as Card. Finally, we also report the absolute value of the smallest non-zero loading denoted as MinLoad, sometime expressed as a percentage contribution, Min PCont = $100 \times |a_{ij}| / \sum_i |a_{ij}|$.

5.1.2 Zou’s synthetic data

Zou et al. (2006) generated an artificial 10×10 covariance matrix with underlying dimension of two starting from three hidden variables:

$$V_1 = N(0, 290), V_2 = N(0, 300), V_3 = -0.3V_1 + 0.925V_2 + \epsilon.$$

where ϵ is a Standard Normal variable. The manifest variables were then generated as

$$X_i = V_1 + \epsilon_i, i = 1, 2, 3, 4; X_i = V_2 + \epsilon_i, i = 5, 6, 7, 8; X_i = V_3 + \epsilon_i, i = 9, 10$$

where the ϵ_i ’s are independent Standard Normal variables. From the correlation matrix among the variables shown in Table 1 it is easy to see that there exist three blocks of variables.

Table 1: Correlations between variables in Zou’s synthetic data.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1	0.996	0.996	0.996	0	0	0	0	-0.3	-0.3
x_2	0.996	1	0.996	0.996	0	0	0	0	-0.3	-0.3
x_3	0.996	0.996	1	0.996	0	0	0	0	-0.3	-0.3
x_4	0.996	0.996	0.996	1	0	0	0	0	-0.3	-0.3
x_5	0	0	0	0	1	0.997	0.997	0.997	0.95	0.95
x_6	0	0	0	0	0.997	1	0.997	0.997	0.95	0.95
x_7	0	0	0	0	0.997	0.997	1	0.997	0.95	0.95
x_8	0	0	0	0	0.997	0.997	0.997	1	0.95	0.95
x_9	-0.3	-0.3	-0.3	-0.3	0.95	0.95	0.95	0.95	1	0.948
x_{10}	-0.3	-0.3	-0.3	-0.3	0.95	0.95	0.95	0.95	0.948	1

Table 2 shows the Varimax rotation of the first three PCA loadings together with the LS-SPCA(BB) solutions with the same cardinality. The last two columns are the LS-CSPCA correlated solutions with cardinality of one. The SPCA methods *SPCA*, *DSPCA*, *SCoTLASS*, *BB-sPCA* and *SPCA-IE* all find the first two sparse solutions that are equal to the varimax rotation of the PCA loadings. These solutions clearly identify the block structure of the data but they are not optimal neither with respect to the variance explained nor to sparsity. In fact, the LS-SPCA solutions of the same cardinality explain efficiently more variance. Furthermore, most of these methods do not require uncorrelatedness of the components. The correlated LS-CSPCA solutions in the same table are much more parsimonious, efficient and interpretable solutions. Of course, on such small matrix the difference in variance explained is relatively small but on larger problems it would be much larger.

Table 2: Varimax rotated PCA loadings, uncorrelated LS-SPCA(BB) loadings with same cardinality and correlated LS-CSPCA(BB) loadings with cardinality (1,1).

	Varimax and SPCA			LS-SPCA-BB		CLS-SPCA-BB (1,1)	
	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 1	Comp 2
X_1		0.5			0.516		
X_2		0.5			0.516		
X_3		0.5			0.516		
X_4		0.5		0.312			1
X_5	0.5						
X_6	0.5						
X_7	0.5			0.536	-0.45		
X_8	0.5			0.536			
X_9			0.71				
X_{10}			0.71	0.572		1	
PVE	58.2	41.3	0.1	60.0	39.6	59.8	39.5
PCVE	58.2	99.4	99.8	60.0	99.6	59.8	99.3
PRCVE	96.9	99.7	100	99.9	99.9	99.5	99.6
Card	4	4	2	4	4	1	1
MinLoad	0.5	0.5	0.71	0.312	0.451	1	1

SPCA methods are designed to explain the most possible variance of the data but sometimes authors speak of block identification and model selection; for example, d’Aspremont et al. (2008) build their method for explaining the most variance with *statistical fidelity* but then venture into applying the sparse components to regression model selection. Then it is not clear why identifying data blocks or explaining an exogenous variable should be a feature of the solutions. In the literature can be found methods for sparse regression, for example.

5.1.3 Pitprops data

The Pitprops dataset was used by Jeffers (1967) to illustrate the difficulty of interpreting PCs and has become a standard benchmark for SPCA methods. It consists of the correlation

matrix of thirteen measures taken on pitprops selected according to a sampling design. Most of the existing SPCA methods were tested on this dataset. The small size of this problem does not allow for extensive comparisons and the differences in variance explained are not very large. This can be appreciated by observing the summary statistics of the distribution of the variance explained by the PCs of first PCs of the principal covariance submatrices and LS-SPCA solutions for all possible combinations of cardinality from four to seven, shown in Figure 1. As expected, for every cardinality a large proportion of LS-SPCA solutions explain more variance than the PCs of the covariance submatrices. However, the maximum variance explained by the PCs is close to that of the LS-SPCA solutions. Hence, small differences are relevant for this example.

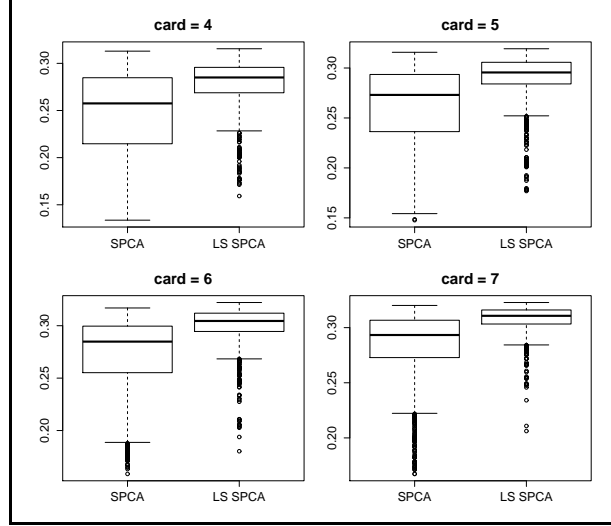


Figure 1: distribution of the variance explained by the first PCs of the principal covariance submatrices and LS-SPCA solutions for all possible combinations of cardinality from four to seven.

Unfortunately, different methods were tested with different cardinalities, so it is impossible to compare each one with the others. Thus, we compare our methods with each one of them using the same cardinalities of the published results. Table 3 compares the results obtained with LS-CSPCA(BB), LS-SPCA(BB) and LS-CSPCA(BE) with those of other SPCA methods found in the literature, which are: SPCA-IE, GSPCA and ESPCA, Gpower, BB-sPCA (two different cardinalities, "7" and "6"), SPCA, DSPCA (two different cardinalities, "6" and "7"), SCoTLASS, DC-PCA and sPCA-rSVD. The results for GSPCA, Gpower and DC-PCA are grouped together under the label "GGD" because they are identical. Some of the LS-SPCA components could not be computed because the cardinality was too low to enforce uncorrelatedness.

Table 3: Pitprops data: comparison of the cumulative variance explained by different methods.

Method	ESPCA	SPCA-IE	GGD	DSPCA6	SCoTLASS	BB-sPCA6	DSPCA7	rSVD	SPCA	BB-sPCA7
Card.	5 2 2	6 2 2	6 2 2	6 2 3	6 6 7 8	6 7 7 8	7 2 3	7 2 4 7	7 4 4 1	7 4 4 1
Comp.										
1st	31.0	31.3	31.3	30.7	30.9	31.3	31.5	31.8	30.4	31.9
2nd	46.9	47.5	47.5	46.4	47.5	48.8	47.4	47.8	46.6	48.9
3rd	59.4	61.0	60.1	59.7	63.5	63.4	60.1	62.8	61.9	57.9
4th					71.0	71.8		71.9	70.2	68.0
LS-CSPCA(BB)										
1st	31.9	32.2	32.2	32.2	32.2	32.2	32.3	32.3	32.3	32.3
2nd	48.3	48.7	48.7	48.7	50.2	50.3	48.7	48.7	49.9	49.9
3rd	60.9	61.3	61.3	62.3	64.5	64.7	62.4	63.0	63.6	63.6
4th					73.2	73.2		71.6	71.6	71.6
LS-SPCA(BB)										
1st	31.9	32.2	32.2	32.2	32.2	32.2	32.3	32.3	32.3	32.3
2nd	48.2	48.4	48.4	48.4	50.2	50.3	48.5	48.5	49.8	49.8
3rd				60.7	64.5	64.7	60.8	62.1	63.4	63.4
4th					73.2	73.2		71.1		
LS-CSPCA(BE)										
1st	31.6	32.0	32.0	32.0	32.0	32.0	32.3	32.3	32.3	32.3
2nd	47.9	48.2	48.2	48.2	49.9	50.1	48.7	48.7	49.8	49.8
3rd	60.5	59.7	59.7	61.1	64.2	64.4	62.3	63.0	63.5	63.5
4th					72.8	73.0		71.6	71.7	71.7

The LS-CSPCA(BB) solutions consistently explain more variance than any of the other methods. The only case in which they explain less variance than an SPCA method is for the fourth dimension of rSVD. The uncorrelated LS-SPCA(BB) components explain less variance than the LS-CSPCA ones, but, still, more than those of the SPCA methods. The BE correlated algorithm perform worse than both BB solutions but also performs well when compared with the other methods, explaining less variance than rSVD, SPCA-IE and the GGD group only with the fourth component. Comparing the variance explained by the first components, it is evident that variance explained by the SPCA methods is never close to the optimum (the LS-SPCA solutions) or even to the suboptimum provided by the BE solutions.

Table 4 compares the variance (L_2 norms) of the first components computed with various SPCA methods with those of the corresponding LS-SPCA solutions. In every case the LS-SPCA components have smaller L_2 norm but explain more variance, thus verifying that components with larger L_2 norm do not necessarily explain the most variance.

Table 4: Pitprops data: L_2 norms of the first components computed with various SPCA methods and with LS-SPCA using the same cardinality.

Method	ESPCA	SPCA-IE	GGD	DSPCA6	SCoTLASS	BB-sPCA6	DSPCA7	rSVD	SPCA	BB-sPCA7
cardinality	5	6	6	6	6	6	7	7	7	7
All methods	3.41	3.74	3.74	3.46	3.71	3.77	3.82	3.99	3.64	4.00
LS-SPCA	2.29	2.78	2.78	2.78	2.78	2.78	3.28	3.28	3.28	3.28

Trendafilov (2013) presents results for the Pitprops data for various new SPCA solutions with yet different cardinalities. These methods optimize different approximations of the correlation matrix under L_1 constraints, discussing which is beyond the scope of this paper. Therefore, we will refer to each method as $Tn\mu = m$, where n refers to the reference number of the formula defining the objective function in the original paper and μ is the value of the tuning parameter used. The T10 method computes uncorrelated components, therefore we used LS-SPCA(BB) for this comparison and LS-CSPCA(BB) for all other cases.

Table 5 compares the summary results of Trendafilov with the corresponding LS-SPCA(BB)

ones. Clearly, also in this case the LS-SPCA components explain more variance, most markedly with the first few ones. The only exception is the second component of the T8 method, which explains slightly more variance than the corresponding LS-CSPCA one. However, the following components perform decidedly worse. Note that, in most cases, Trendafilov’s solutions present small loadings in the first components.

Table 5: Pitprops data: Trendafilov’s solutions compared with the corresponding LS-CSPCA(BB) ones.

Component	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6
Method	T5 $\mu = 5$, Card (10 6 3 1 3 2)						LS-CSPCA(BB)					
PRCVE	99.3	93	95.7	95.7	95.3	96.1	100	99.3	97.6	97.1	97.3	97.9
MinLoad	0.017	0.2	0.322	1	0.024	0.272	0.110	0.147	0.458	1	0.415	0.550
Method	T7 $\mu = 6$, Card (7 2 3 1 1 1)						LS-CSPCA(BB)					
PRCVE	95.7	91.9	91.1	90.1	92.2	92.9	99.5	96.1	95.7	96	96.1	96.7
MinLoad	0.059	0.707	0.122	1	1	1	0.289	0.467	0.418	1	1	1
Method	T8 $\mu = 3$, Card (7 1 3 2 1 1)						LS-CSPCA(BB)					
PRCVE	92.5	82	77.5	87.6	89.8	91	99.5	81.5	89	95.5	94.3	95.5
MinLoad	0.081	1	0.432	0.707	1	1	0.289	1	0.418	0	1	1
Method	T9 $\mu = 4$, Card (7 2 1 1 1 1)						LS-CSPCA(BB)					
PRCVE	97.3	92.3	84.7	84.7	89.4	92.4	99.5	96.1	87.7	91.4	94.2	94.9
MinLoad	0.102	0.708	1	1	1	1	0.289	0.467	1	1	1	1
Method	T10 $\mu = 5$, Card (7 6 4 4 5 6)						LS-SPCA(BB)					
PRCVE	97.6	92.6	96.6	96.7	97.4	97.5	99.5	99	97.9	98	98	98.3
MinLoad	0.132	0.186	0.146	0.059	0.031	0.063	0.289	0.152	0.320	0.14296	0.137	0.048
Method	T11 $\mu = 4$, Card (7 4 1 1 2 3)						LS-CSPCA(BB)					
PRCVE	96.3	93.9	86	90.2	92.8	93.6	99.5	98.4	91.7	92	96.1	97
MinLoad	0.016	0.059	1	1	0.087	0.091	0.289	0.234	1	1	0.530	0.434

5.2 Examples

In this section we illustrate some results obtained by applying LS-SPCA to datasets of different dimensionality available in the literature or on the StatLib Data Archive (*lib.stat.cmu.edu*) and UCI Machine Learning Repository (*archive.ics.uci.edu*, Frank and Asuncion, 2010). Since there is a trade-off between explaining the variance and sparsity, the results are not necessarily the "best" ones with respect to either requirements. We present what we obtained with what we consider reasonable working requirements for illustrative purposes.

Anthropometric measures

The oldest application of PCA is the analysis of a set of seven anthropometric measures taken on a sample of non-habitual criminals (Macdonell, 1902). The data were used to classify criminals by their physical features. We use this small dataset to illustrate the differences among the LS-SPCA solutions obtained with the BB and BE algorithms and the globally optimal ones. We obtained the global optima by exploring all possible solutions for three components with cardinality (2, 2, 3) and (2, 3, 3). The summary results are shown in Table 6. As expected, the first BB components explain more total variance of all methods but the complete search attains a higher global optimum. The slightly higher PCVE of the first solution results in a loss of PCVE when other components are included; for cardinalities (2, 2, 3) the two BB components still perform better than the global optimum, but not for cardinalities (2, 3, 3). The BE algorithm performs notably worse than the BB one for the cardinality (2, 2, 3) but in the other case it outperforms it by finding the optimal solutions.

In this case the BB search pays for its initial greediness by getting stuck in a local maximum while BE does not.

This is the only example we observed in which BE outperforms BB in early components, other cases were observed for a larger number of components and by small differences. In our observations, the BE solutions are usually not much worse than the BB ones.

Table 6: Anthropometric measures: summary results of the globally optimal, BB and BE solutions with cardinalities (2, 2, 3) and (2, 3, 3).

Method	Optimal			BB			BE		
	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3	Comp 1	Comp 2	Comp 3
Three components, cardinality 2 2 3									
PVE	49.2	13.6	18.7	49.3	18.7	10.9	49.2	11.6	12.6
PCVE	49.2	62.7	81.5	49.3	67.9	78.8	49.2	60.8	73.4
PRCVE	90.6	82.9	95.8	90.8	89.7	92.7	90.6	80.3	86.3
Min PCont	44.4	34.2	15.4	0	29.7	26.1	44.4	34.1	10.1
Three components, cardinality 2 3 3									
PVE	49.2	23.2	10.7	49.3	20.4	11.2	49.2	23.2	10.7
PCVE	49.2	72.3	83	49.3	69.7	80.9	49.2	72.3	83
PRCVE	90.6	95.5	97.6	90.8	92	95.1	90.6	95.5	97.6
Min PCont	44.4	20.8	12.8	0	29.7	27.5	44.4	20.8	12.8

The results in this example can be replicated using the Anthro dataset included in the R package *spca*.

Baseball hitters

This dataset (available at Statlib) contains observations on 16 performance statistics of 263 US Major League baseball hitters taken on their career and on 1986 season. The data were used to explain the players' salary.

We computed five components with the BE algorithm, trimming the percent contributions to the thresholds 0.35, 0.35, 0.2, 0.2 and 0.2, respectively. All the components but the fourth one could be trimmed to the required threshold; for the fourth one trimming ended because the minimal cardinality to permit uncorrelatedness was reached. The summary results are shown in Table 7 together with those of the BB solutions computed with the same cardinalities. The latter method explains slightly more variance but the last two components present a percent contribution of three percent or less.

Table 7: Baseball Hitters data: comparison of LS-SPCA(BB) and LS-SPCA(BE) results.

Statistics	BE					BB				
	Comp1	Comp2	Comp3	Comp4	Comp5	Comp1	Comp2	Comp3	Comp4	Comp5
PVE	44.5	24.6	10.9	5.7	4.1	44.5	24.7	10.8	5.7	4.4
PCVE	44.5	69.1	79.9	85.6	89.7	44.5	69.2	80.1	85.7	90.1
PRCVE	98.2	97.3	97.7	98.1	98	98.2	97.5	97.9	98.3	98.3
Card	3	3	4	4	7	3	3	4	4	7
MinPContr	24.8	27.5	13.4	8.1	11.1	24.5	18.0	14.4	3.0	1.7

The results in this example can be replicated using the Anthro dataset included in the R package *spca*.

Optical Recognition of Handwritten Digits

The optdigits dataset (available at the UCI Repository) contains measures on graphical attributes of different handwritten digits, which were used to classify the digits. We merged

the training and test samples and removed the classification variable and two constant ones, which left 62 variables. We ran the BE algorithm requiring that the variance explained by each component was at least 90% of that explained by the initial untrimmed solution (with reference to Algorithm 1, we set the *mvl* threshold to 0.1). Table 8 shows the summary results for the first five components. Each solution gives over 91% of PRCVE with at most 11 loadings.

Table 8: Optical Recognition data: LS-SPCA(BE) summary results.

	Comp1	Comp2	Comp3	Comp4	Comp5
PVE	10.7	9.2	7.1	5.4	4.5
PCVE	10.7	19.9	27.0	32.4	36.9
PRCVE	92.1	91.3	91.3	91.5	91.7
Card	6	7	7	10	11
Min PContr	14.2	9.8	10.3	7.4	6.0

US crime data

This dataset (available at the UCI repository) contains socioeconomic records on different US cities collected in the 90s. The data was used to explain the rate of violent crime in each city. We deleted 22 variables and one observation with missing values. The final set contains 1994 observations on 99 variables. We run LS-SPCA(BE) requiring that the cumulative variance explained after including each component was at least 95% of that explained by the PCs. The summary results of first five components are shown in Table 9. The solutions explain over 95% of the variance explained by the ordinary PCs with extremely low cardinality and all contributions above 11%.

Table 9: US crime data: LS-SPCA(BE) summary results.

	Comp1	Comp2	Comp3	Comp4	Comp5
PVE	24.1	16.4	8.7	7.3	5.6
PCVE	24.1	40.5	49.2	56.5	62.1
PRCVE	95.4	95.8	95.3	95.4	95.6
Card	2	4	3	6	6
MinPContr	40.6	13.1	25.3	11.7	11.3

One hundred plant species leaves dataset

The *100 leaves plant species* dataset (available at the UCI repository) was obtained by merging the measurements on three different aspects of the shapes of leaves of one hundred different species of plants. It was used for classifying the plants. We removed one observation from two files because it was missing from the other. So, the final set contained 1599 instances of 186 variables. We ran LS-SPCA(BE) requiring that each of the first five component computed explained at least 90% of the variance explained by the initial untrimmed solution. The summary results are shown in Table 10. The PRCVE for first component is 98% with just two loadings, the addition of the second component leads to a PRCVE of almost 96%, with a total of just seven loadings, and for the first three PRCVE is 95.9%, with just a total of 11 loadings. The following components have higher cardinality and also explain over 95% of the variance explained by the PCs.

Table 10: Plant Leaves data: LS-SPCA(BE) summary results.

	Comp1	Comp2	Comp3	Comp4	Comp5
PPVE	23.7%	9.5%	6.0%	5.0%	3.8%
PCVE	23.7%	33.3%	39.3%	44.3%	48.1%
PRCVE	97.9%	95.9%	95.4%	95.3%	95.1%
Card	2	5	4	10	12
Min PContr	49.1	7.2	16.6	7.1	5.5

Isolated Letter Speech Recognition

The *isolet1+2+3+4* dataset (available at the UCI Repository) contains 617 measurements taken on the sounds produced by readers speaking the name of different letters of the alphabet. The first eight PCs explain 50.4% of the total variance. Therefore, we ran LS-SPCA(BE) requiring not more than 10 components or up to 50% of total variance explained. For the first three components trimming was stopped if the loss of PVE from the initial component was greater than 10%, for the next three components this value was 15% and 20% for the remaining ones. The algorithm took 158 minutes to terminate with 10 components and 48.8% total variance explained. The summary results are shown in Table 11. The first component explains 93.7% of the variance explained by the first PC with just three non-zero loadings. The addition of the following components leads to explaining decreasing proportions of the variance explained by the PCs, but always around 90% of it. The loadings have at most cardinality of ten and only one is smaller than 0.1.

Table 11: Speech recognition data: LS-SPCA(BE) results.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10
PVE	18.1	8.2	5.0	3.9	3.6	2.5	2.3	1.9	1.8	1.6
PCVE	18.1	26.2	31.2	35.1	38.7	41.2	43.5	45.4	47.2	48.8
PRCVE	93.7	93.1	92.8	92.1	91.9	90.9	90.6	90.0	89.8	89.5
Card	3	10	6	9	10	8	9	9	9	10
MinLoad	0.438	0.233	0.303	0.256	0.227	0.092	0.210	0.126	0.192	0.238

Summary of the results

Table 12 shows the cumulative variance explained by the LS-SPCA(BE) components and their cardinalities. For all datasets there is a very consistent reduction of the cardinality while a large proportion of the variance explained by the PCs is preserved. As mentioned above, there is a trade-off between low cardinality and high proportion of variance explained, therefore these results are only indicative of a generic performance of the LS-SPCA(BE) algorithm and can be improved in favor of either feature.

Table 12: PVE and cardinality of the LS-SPCA(BE) results for the different datasets considered.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10
Dataset and dimension	Baseball Hitters, $p = 13$									
PRCVE	98.2	97.5	97.9	98.3	98.2					
Cardinality	3	3	4	4	5					
Dataset and dimension	Optical Recognition, $p = 62$									
PRCVE	92.1	91.3	91.3	91.5	91.4	92.1	92.1	92.2	92.7	93.1
Cardinality	6	7	7	10	10	13	12	12	13	15
Dataset and dimension	US Crime, $p = 99$									
PRCVE	95.4	95.8	95.3	95.4	95.6	95.5	95.6	95.3		
Cardinality	2	4	3	6	6	8	8	9		
Dataset and dimension	100 Leaves, $p = 186$									
PRCVE	97.9	95.9	95.4	95.3	95.1	94.6	94.3	94.4	94.4	94.2
Cardinality	2	5	4	10	12	6	14	20	13	15
Dataset and dimension	Isolated Letters, $p = 617$									
PRCVE	93.7	93.1	92.8	92.1	91.9	90.9	90.6	90.0	89.8	89.5
Cardinality	3	10	6	9	10	8	9	9	9	10

5.2.1 Computational times

We present a study on the computational time taken by LS-SPCA(BB) and LS-SPCA(BE) algorithms. The elapsed time is computed on a 64bit quadri-core Intel i5[®] CPU with and 4Gb RAM, using non-optimized R (R Core Team, 2013) code.

Branch-and-bound searches are known to take exponential time; Farcomeni (2009) shows how the BB algorithm becomes very time consuming as the dimension of the matrices increases. A comparison of the computing times taken by the LS-SPCA(BB) and LS-SPCA(BE) to compute solutions of increasing complexity on the Pitprops data is shown in Table 13. Clearly the BE algorithm is much faster with a peak of 141 times shorter computational time. These results should be evaluated considering that the complexity for the BE algorithm decreases as the cardinality increases while for the BB one it is maximal when the cardinality is near half the dimension of the problem.

Table 13: Pitprops data: comparison of the computational times taken by the BE and BB algorithms for different cardinalities. Time in seconds.

Cardinality	Replications	BE	BB	Relative
2 2 3 4 5	30	4.43	128.17	28.932
2 2 6 6 6	30	2.97	147.5	49.663
6 6 3 4 5	30	4.49	259.44	57.782
2 6 6 6 6	30	3.78	289.42	76.566
6 6 6 6 6	30	1.95	275.41	141.236

Table 14 shows a comparison of the computational times taken by the BB and BE algorithms to compute 5 components of cardinality 10 on datasets of increasing dimension. The computational times increase exponentially with the dimension of the dataset.

Table 14: Different datasets: comparison of LS-SPCA(BE) computational times for different number of dimensions. Time in seconds.

Dataset	Dimension	Replications	Average	Relative	Sec per dimension
Optical Digits	62	100	1.1	1	0.02
Crime in US	99	100	4.7	4.2	0.05
100 Leaves	186	100	40.9	36.8	0.22
Isolated Letters	617	20	6766.4	6079.4	10.97

Finally, we compared the the computational time for trimming a different number of

loadings at each iteration. The results of computing 5 components of cardinality 10 on the Isolated Letters dataset trimming 1, 5, 10, 20 and 50 loadings at the time are shown in Table 15. As expected, the computational time decreases as the number of loading trimmed increases; the relationship, for this example, is of log-log type.

Table 15: Time taken to compute 5 components on the isolet dataset for increasing number of loadings trimmed at each iteration. Time in minutes.

Trimmed	Replications	Average Time	Relative
1	20	112.773	36.4
5	20	24.8	8.0
10	20	11.4	3.7
20	20	5.6	1.8
50	20	3.1	1

6 Discussion

The popularity of PCA is due to its ability of summarising a set of variables with few components. SPCA greatly enhances the interpretability of these components. SPCA-LS represents an improvement over other SPCA methods because it maintaining the PCs' key properties: uncorrelatedness and LS optimality. The model based approach adopted overcomes the difficulties in defining deflations and variance explained existing for other methods.

The problem of simplifying the components has been widely discussed in the Factor Analysis literature. From the discussion it comes out clearly that there are different definitions and requirements for "simplicity", (for example, the ones suggested by Thurstone, 1947), which cannot all be included in an objective function. In fact solving the PCA problem with restrictions on the cardinality achieves sparsity but does not guarantee large loadings or efficiency in explaining the variance. In this sense, the BE algorithm can be a useful tool for including more simplicity criteria into the problem by constraining the way the solutions are found. In our implementation of the BE algorithm we include the possibility of excluding from a solution variables that are already in previous ones (so achieving parsimony) and the possibility of selecting in advance we variables can enter a component. Hence, the BE algorithm is very useful for selecting a "good" solution, that is one that is easy to interpret but that satisfies other requirements. This is very difficult to achieve with black-box numerical solvers, for as sophisticated and scalable as they could. Instead, using adaptable searches requirement like identifying blocks of data could be easily met. However, it could be useful to develop a more efficient computational algorithm for finding the LS-SPCA solutions on larger problems, of the kind existing for the maximisation of the variance of the components.

References

- Cadima, J. and Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22(2):203–214.
- d'Aspremont, A., Bach, F., and Ghaoui, L. E. (2008). Optimal Solutions for Sparse Principal Component Analysis. *Journal of Machine Learning Research*, 9:1269–1294.
- d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448.
- Farcomeni, A. (2009). An exact approach to sparse principal component analysis. *Computational Statistics*, 24(4):583–604.

- Frank, A. and Asuncion, A. (2010). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. <http://archive.ics.uci.edu/ml>.
- Hotelling, H. (1933). Analysis of a Complex of Statistical Variables with Principal Components. *Journal of Educational Psychology*, 24:498–520.
- Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.
- Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer New York.
- Jeffers, J. (1967). Two case studies in the application of principal component. *Applied Statistics*, 16:225–236.
- Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):pp. 531–547.
- Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553.
- Macdonell, W. (1902). Criminal Anthropometry and the Identification of Criminals. *Biometrika*, 1(2):177–227.
- Mackey, L. (2009). Deflation methods for sparse pca. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1017–1024.
- Magnus, J. and Neudecker, H. (1999). *Matrix Differential Calculus With Applications in Statistics and Econometrics*. John Wiley.
- Moghaddam, B., Weiss, Y., and Avidan, S. (2006). Spectral bounds for sparse pca: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, pages 915–922. MIT Press.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, C. and Toutenburg, H. (1999). *Linear Models: LS and Alternatives*. Springer Series in Statistics Series. Springer Verlag, second edition.
- Shen, H. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 101:10151034.
- Sriperumbudur, B., Torres, D., and Lanckriet, G. (2009). A D.C. programming approach to the sparse generalized eigenvalue problem. *Computer Engineering*, 1(1):40.
- ten Berge, J. (1993). *Least Squares Optimization in Multivariate Analysis*. DSWO Press, Leiden University.
- Thurstone, L. (1947). *Multiple-factor analysis*. The University of Chicago Press.
- Trendafilov, N. (2013). From simple structure to sparse components: a review. *Computational Statistics*, 28(4).

- Trendafilov, N. and Jolliffe, I. (2006). Projected gradient approach to the numerical solution of the scotlass. *Comput. Stat. Data Anal.*, 50(1):242–253.
- Wang, Y. and Wu, Q. (2012). Sparse pca by iterative elimination algorithm. *Advances in Computational Mathematics*, 36:137–151.
- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In Krishnaiah, P., editor, *In Multivariate Analysis*, volume 59, pages 391–420. Academic Press, NY.
- Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Appendix

Proof of the result

The extra sum of squares principle says that the variance explained by an additional variable in a regression model is equal to variance explained by its orthogonal complement to the other variables. Consequently, the variance explained by a component computed under uncorrelatedness constraints is a sup for the Vexp of correlated components.

Proposition 1 (extra sum of squares principle). *A component correlated with the preceding ones cannot explain more variance than its complement orthogonal to the others.*

Proof. Let $\mathbf{T} = \mathbf{XA}$ be a set of components and $\mathbf{z} = \mathbf{Xb}$ be another component such that $\mathbf{z}'\mathbf{t}_j \neq 0$ for at least one j . Also let $\mathbf{P} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'$ be the orthogonal projector onto the space of \mathbf{T} and $\mathbf{Q} = \mathbf{I} - \mathbf{P}$ its complement.

The variance explained by the components \mathbf{T} is $\text{tr}(\mathbf{X}'\mathbf{P}\mathbf{X})$. We can write the components $\mathbf{G} = [\mathbf{T}; \mathbf{z}]$ as $\mathbf{G} = \mathbf{P}\mathbf{G} + \mathbf{Q}\mathbf{G} = \mathbf{TM} + \tilde{\mathbf{z}}$, where \mathbf{M} is a matrix and $\tilde{\mathbf{z}} = \mathbf{Q}\mathbf{z}$ is the complement of \mathbf{z} orthogonal to the \mathbf{T} components. Then, the regression of \mathbf{X} onto \mathbf{G} is $\hat{\mathbf{X}} = \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}\mathbf{X} = \mathbf{P}\mathbf{X} + \tilde{\mathbf{z}}(\tilde{\mathbf{z}}'\tilde{\mathbf{z}})^{-1}\tilde{\mathbf{z}}'\mathbf{X}$. It follows that the variance explained by the components in \mathbf{G} is given by $\text{tr}(\hat{\mathbf{X}}'\hat{\mathbf{X}}) = \text{tr}(\mathbf{X}'\mathbf{P}\mathbf{X}) + \text{tr}(\mathbf{X}'\tilde{\mathbf{z}}(\tilde{\mathbf{z}}'\tilde{\mathbf{z}})^{-1}\tilde{\mathbf{z}}'\mathbf{X})$. Therefore, the variance explained by the correlated component \mathbf{z} is equal to that explained by its orthogonal complement to the \mathbf{T} variables, completing the proof. \square

Uncorrelated LS-SPCA components

Here we derive the solutions for Problem (8) based on a generalization of the proof for constrained multiple regression as given in (Magnus and Neudecker, 1999, Theorem 13.5). We adopt the notation defined in Section 2, assuming that \mathbf{D}_j is invertible and that $\mathbf{R}_j'\mathbf{R}_j$ is singular.

Proposition 2. *The uncorrelated sparse principal components of given indices ind_j that minimize the L_2 norm of the residuals of the approximation are defined as*

$$\mathbf{a}_j = \arg \min_{\mathbf{a}_j \in \mathbb{R}^p} \|\mathbf{X} - \mathbf{X}\mathbf{a}_j\mathbf{p}_j'\| = \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \frac{\mathbf{a}_j'\mathbf{S}\mathbf{S}\mathbf{a}_j}{\mathbf{a}_j'\mathbf{S}\mathbf{a}_j}, \quad j = 1, \dots, d \quad (15)$$

subject to $a_{i,j} = 0$ if $i \notin \text{ind}_j$ and $\mathbf{R}_j\mathbf{a}_k = \mathbf{0}$, $j < k$.

The solutions are given by the eigenvectors satisfying

$$\left[\mathbf{I}_{n_j} - \mathbf{D}_j^{-1}\mathbf{R}_j'(\mathbf{R}_j\mathbf{D}_j^{-1}\mathbf{R}_j')^+\mathbf{R}_j \right] \mathbf{B}_j^*\tilde{\mathbf{a}}_j = \mathbf{C}_j\mathbf{B}_j^*\tilde{\mathbf{a}}_j = \phi_{\max}\tilde{\mathbf{a}}_j, \quad (16)$$

where the subscript “+” denotes a generalized inverse, $\mathbf{C}_j = [\mathbf{I}_{n_j} - \mathbf{D}_j^{-1}\mathbf{R}_j'(\mathbf{R}_j\mathbf{D}_j^{-1}\mathbf{R}_j')^+\mathbf{R}_j]$, $\mathbf{C}_1 = \mathbf{I}_{n_1}$, and $\mathbf{B}_j^* = \mathbf{D}_j^{-1}\mathbf{J}_j'\mathbf{S}\mathbf{S}\mathbf{J}_j$. The solutions exist because \mathbf{R}_j' spans the space of \mathbf{W}_j' .

Proof. The objective function can be easily derived from the general problem in Equation (4) by first observing that the \mathbf{P}' matrix is a matrix of regression coefficients for the components $\mathbf{T} = \mathbf{XA}$. By developing the L_2 norm as a trace and using the uncorrelatedness of the components it is easy to obtain

$$\mathbf{A} = \arg \min_{\mathbf{a}_j \in \mathbb{R}^{p \times d}} \|\mathbf{X} - \mathbf{XAP}'\| = \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \sum_{j=1}^d \frac{\mathbf{a}_j' \mathbf{S} \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j},$$

subject to $a_{i,j} = 0$ if $i \notin \text{ind}_j$ and $\mathbf{t}_j' \mathbf{t}_k = 0$, $j \neq k$.

Because of the uncorrelatedness the problem can be solved separately for each component as in Equation (15).

By Problem (15), we want to maximize the variance explained subject to $\mathbf{R}_j \mathbf{J}_j \tilde{\mathbf{a}}_j = 0$. Hence, we need to maximize the Langrangian:

$$\begin{aligned} L(\tilde{\mathbf{a}}_j, \boldsymbol{\lambda}) &= \text{tr} [\mathbf{X}' \mathbf{t}_j (\mathbf{t}_j' \mathbf{t}_j)^{-1} \mathbf{t}_j' \mathbf{X}] - 2\boldsymbol{\lambda}' \mathbf{R}_j \tilde{\mathbf{a}}_j \\ &= (\tilde{\mathbf{a}}_j' \mathbf{D}_j \tilde{\mathbf{a}}_j)^{-1} \tilde{\mathbf{a}}_j' \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j - 2\boldsymbol{\lambda}' \mathbf{R}_j \tilde{\mathbf{a}}_j. \end{aligned} \quad (17)$$

Equating the partial derivatives to zero gives:

$$\frac{\partial L}{\partial \tilde{\mathbf{a}}} = -\mathbf{D}_j \tilde{\mathbf{a}}_j \alpha_j + \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j \beta_j - \mathbf{R}_j' \boldsymbol{\lambda} = \mathbf{0} \quad (18)$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \mathbf{R}_j \tilde{\mathbf{a}}_j = \mathbf{0}, \quad (19)$$

where $\alpha_j = \tilde{\mathbf{a}}_j' \mathbf{J}_j' \mathbf{S}_j \mathbf{S}_j \mathbf{J}_j \tilde{\mathbf{a}}_j (\tilde{\mathbf{a}}_j' \mathbf{D}_j \tilde{\mathbf{a}}_j)^{-2}$ and $\beta_j = (\tilde{\mathbf{a}}_j' \mathbf{D}_j \tilde{\mathbf{a}}_j)^{-1}$. Since for the first component $\mathbf{R}_1 = \mathbf{0}$, the solution is

$$\mathbf{D}_1^{-1} \mathbf{J}_1' \mathbf{S} \mathbf{S} \mathbf{J}_1 \tilde{\mathbf{a}}_1 = \tilde{\mathbf{a}}_1 \frac{\alpha_j}{\beta_j}. \quad (20)$$

For the subsequent components we have that premultiplying equation (18) by \mathbf{D}_j^{-1} gives

$$-\tilde{\mathbf{a}}_j \alpha_j + \mathbf{D}_j^{-1} \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j \beta_j - \mathbf{D}_j^{-1} \mathbf{R}_j' \boldsymbol{\lambda} = \mathbf{0} \quad (21)$$

Premultiplying the above by \mathbf{R}_j gives

$$\mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j \beta_j = \mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{R}_j' \boldsymbol{\lambda},$$

because of Equality (19). Therefore,

$$\boldsymbol{\lambda} = (\mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{R}_j')^+ \mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j \beta_j.$$

Substituting this in Equation (21) gives

$$\mathbf{D}_j^{-1} \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j - \mathbf{D}_j^{-1} \mathbf{R}_j' (\mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{R}_j')^+ \mathbf{R}_j \mathbf{D}_j^{-1} \mathbf{J}_j' \mathbf{S} \mathbf{S} \mathbf{J}_j \tilde{\mathbf{a}}_j = \tilde{\mathbf{a}}_j \frac{\alpha_j}{\beta_j}.$$

Since we want to maximize the function, the solution is the eigenvector corresponding to the maximum eigenvalue, which is equal to $\frac{\alpha_j}{\beta_j} = \frac{\mathbf{a}_j' \mathbf{S} \mathbf{S} \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}$, as required, completing the proof. \square

Correlated LS-SPCA components

The approximate solutions are computed by requiring that each component gives the LS approximation of the residuals of \mathbf{X} orthogonal to the previously computed ones, $\mathbf{X}_j = \mathbf{Q}_j \mathbf{X}$, with $\mathbf{X}_1 = \mathbf{X}$. Then, the loadings \mathbf{a}_j must satisfy:

$$\mathbf{a}_j = \arg \min_{\mathbf{a}_j \in \mathbb{R}^p} \|\mathbf{X}_j - \mathbf{X} \mathbf{a}_j \mathbf{p}_j'\| = \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \frac{\mathbf{a}_j' \mathbf{S}_j \mathbf{S}_j \mathbf{a}_j}{\mathbf{a}_j' \mathbf{S} \mathbf{a}_j}, \quad j = 1, \dots, d, \quad (22)$$

subject to $a_{i,j} = 0$ if $i \notin \text{ind}_j$.

The components following the first one do not maximise the variance explained. The correlated sparse components can be computed from the covariance matrix. If we let $\mathbf{Z}_j = (\mathbf{I}_p - \mathbf{A}_{(j)}(\mathbf{A}'_{(j)}\mathbf{S}\mathbf{A}_{(j)})^{-1}\mathbf{A}'_{(j)}\mathbf{S})$, $\mathbf{Z}_1 = \mathbf{I}_p$, the residual covariance matrix can be written as $\mathbf{X}'_j\mathbf{X}_j = \mathbf{S}_j = \mathbf{S}\mathbf{Z}_j$. Then, the solutions are the eigenvectors satisfying:

$$\mathbf{D}_j^{-1}\mathbf{J}'_j\mathbf{Z}'_j\mathbf{S}\mathbf{S}\mathbf{Z}_j\mathbf{J}_j\tilde{\mathbf{a}}_j = \phi_{\max}\tilde{\mathbf{a}}_j. \quad (23)$$

Obviously, the first component is the same as the uncorrelated one. The following solutions can be computed from the leftmost eigenvectors, \mathbf{b}_j , of the matrices $\mathbf{D}_j^{-1/2}\mathbf{J}_j\mathbf{Z}'_j\mathbf{S}\mathbf{S}\mathbf{Z}_j\mathbf{J}_j\mathbf{D}_j^{-1/2}$ as $\tilde{\mathbf{a}}_j = \mathbf{D}_j^{-1/2}\mathbf{b}_j$.

Correlated LS-SPCA components with orthogonal loadings

Proposition 3. *The correlated sparse principal components of given indices ind_j that minimize the L_2 norm of the residuals of the approximation subject to orthogonality constraints are defined as*

$$\begin{aligned} \mathbf{a}_j &= \arg \min_{\mathbf{a}_j \in \mathbb{R}^p} \|\mathbf{X}_j - \mathbf{X}\mathbf{a}_j\mathbf{p}'_j\| = \arg \max_{\mathbf{a}_j \in \mathbb{R}^p} \frac{\mathbf{a}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{a}_j}{\mathbf{a}'_j\mathbf{S}\mathbf{a}_j}, \quad j = 1, \dots, d, \\ &\text{subject to } a_{i,j} = 0 \text{ if } i \notin ind_j \text{ and } \mathbf{a}'_j\mathbf{A}_j = \mathbf{0} \end{aligned} \quad (24)$$

The first solution is the same as for the uncorrelated case. The following ones are given by the eigenvectors satisfying

$$\mathbf{D}_j^{-1}\mathbf{J}'_j \left[\mathbf{I}_p - \mathbf{A}_j(\mathbf{A}'_j\mathbf{J}_j\mathbf{D}_j^{-1}\mathbf{J}'_j\mathbf{A}_j)^{-1}\mathbf{A}'_j\mathbf{J}_j\mathbf{D}_j^{-1}\mathbf{J}'_j \right] \mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j = \phi_{\max}\tilde{\mathbf{a}}_j, \quad (25)$$

where ϕ_{\max} is the largest eigenvalue.

Proof. Let $\mathbf{t}_j = \mathbf{W}_j\tilde{\mathbf{a}}_j = \mathbf{X}\mathbf{J}_j\tilde{\mathbf{a}}_j$ be the j -th component. We need to maximize the variance explained subject to $\mathbf{A}'_j\mathbf{J}_j\tilde{\mathbf{a}}_j = \mathbf{0}$. Hence we need to maximize the Lagrangian:

$$\begin{aligned} L(\tilde{\mathbf{a}}_j, \boldsymbol{\lambda}) &= \text{tr} [\mathbf{X}'_j\mathbf{t}(\mathbf{t}'_j\mathbf{t}_j)^{-1}\mathbf{t}'_j\mathbf{X}_j] - 2\tilde{\mathbf{a}}'_j\mathbf{J}'_j\mathbf{A}_j\boldsymbol{\lambda} \\ &= \tilde{\mathbf{a}}'_j\mathbf{J}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j(\tilde{\mathbf{a}}'_j\mathbf{D}_j\tilde{\mathbf{a}}_j)^{-1} - 2\tilde{\mathbf{a}}'_j\mathbf{J}'_j\mathbf{A}_j\boldsymbol{\lambda} \end{aligned} \quad (26)$$

Equating the partial derivatives to zero gives:

$$\frac{\partial L}{\partial \tilde{\mathbf{a}}} = -\mathbf{D}_j\tilde{\mathbf{a}}_j\alpha_j + \mathbf{J}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j\beta_j - \mathbf{J}'_j\mathbf{A}_j\boldsymbol{\lambda} = \mathbf{0} \quad (27)$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \tilde{\mathbf{a}}'_j\mathbf{A}_j = \mathbf{0} \quad (28)$$

where $\alpha_j = \tilde{\mathbf{a}}'_j\mathbf{J}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j(\tilde{\mathbf{a}}'_j\mathbf{D}_j\tilde{\mathbf{a}}_j)^{-2}$ and $\beta_j = (\tilde{\mathbf{a}}'_j\mathbf{D}_j\tilde{\mathbf{a}}_j)^{-1}$. Premultiplying Equation (27) by \mathbf{D}_j^{-1} gives

$$\tilde{\mathbf{a}}_j\alpha_j = \mathbf{D}_j^{-1}\mathbf{J}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j\beta_j - \mathbf{D}_j^{-1}\mathbf{A}_j\boldsymbol{\lambda} \quad (29)$$

Premultiplying the above by \mathbf{A}'_j gives

$$\mathbf{A}'_j\mathbf{D}_j^{-1}\mathbf{J}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j\beta_j = \mathbf{A}'_j\mathbf{D}_j^{-1}\mathbf{A}_j\boldsymbol{\lambda}$$

because of Equality (28). Therefore,

$$\boldsymbol{\lambda} = (\mathbf{A}'_j\mathbf{D}_j^{-1}\mathbf{A}_j)^{-1}\mathbf{A}'_j\mathbf{D}_j^{-1}\mathbf{J}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j\beta_j.$$

Substituting this into Equation (29) gives

$$\mathbf{D}_j^{-1}\mathbf{J}'_j \left[\mathbf{I}_p - \mathbf{A}_j(\mathbf{A}'_j\mathbf{J}_j\mathbf{D}_j^{-1}\mathbf{J}'_j\mathbf{A}_j)^{-1}\mathbf{A}'_j\mathbf{J}_j\mathbf{D}_j^{-1}\mathbf{J}'_j \right] \mathbf{S}_j\mathbf{S}_j\mathbf{J}_j\tilde{\mathbf{a}}_j = \frac{\alpha_j}{\beta_j}\tilde{\mathbf{a}}_j.$$

Since we want to maximize the function, the solution is the eigenvector corresponding to the maximum eigenvalue, which is equal to $\frac{\alpha_j}{\beta_j} = \frac{\mathbf{a}'_j\mathbf{S}_j\mathbf{S}_j\mathbf{a}_j}{\mathbf{a}'_j\mathbf{S}\mathbf{a}_j}$, as required in Equation (24), completing the proof. \square